

# The threshold model as a general purpose normalizing transformation

DEREK A. ROFF

*Department of Biology, McGill University, 1205 Dr Penfield Ave., Montreal, Quebec, Canada, H3A 1B1*

The estimation of heritabilities and genetic correlations is based on the assumption that the trait distributions are normal. When the distributions are not normal it is advisable to transform the data to produce normality. However, it is possible that no suitable transformation can be found. The purpose of the present paper is to point out that the threshold model of quantitative genetics can be used as a generalized transformation. To utilize this method it is only necessary to divide the data at the median (approximately) and code the two halves as 0 and 1. Estimates can then be made using algorithms outlined herein. A simulation study shows that the threshold transformation gave unbiased estimates of the heritability and genetic correlation in all cases. The 95% confidence limits correctly included the true heritability value in the required 95% of cases, while the estimated confidence region for the genetic correlation was also correct provided that the geometric mean heritability was greater than approximately 0.15, a restriction that applied also to the normally distributed data. Confidence intervals estimated from the non-normal data were consistently too small. The method is illustrated using data on the proportion of diapausing eggs produced by the cricket, *Allonemobius socius*.

**Keywords:** genetic correlation, heritability, threshold, transformation.

## Introduction

A fundamental assumption in the estimation of genetic parameters is that the underlying trait distributions are normal (Bulmer, 1985). Unfortunately it is frequently found that the distributions are not normal. There are typically two alternative solutions: first, the lack of normality can be ignored, and second a transformation can be sought that normalizes the data (Falconer, 1989). The former approach is problematical because it is not clear to what extent non-normality can bias estimates of heritability or genetic correlation. However, as noted by Falconer (1989; pp 297–298), ‘The first purpose of experimental observations is the description of the genetic properties of the population, and a scale transformation obscures rather than illuminates the description.’ Nevertheless, when a transformation exists that does normalize the data it would seem wise to estimate the parameters either on the transformed scale alone or on both scales to assess the possible bias introduced by the lack of normality. If the two sets of estimates are found to be very similar it may be more ‘illuminating’ to use the untransformed rather than the transformed data. Unfortunately there may not exist

any simple transformation, such as the logarithmic, that normalizes the data. The purpose of the present paper is to point out that the threshold model of quantitative genetics itself represents a general normalizing transformation.

To make use of the threshold model we need simply assume that there exists some transformation that will normalize the data. Given that such a transformation exists, we can estimate the heritability of the transformed data even without actually doing the transformation by making use of the threshold model of quantitative genetics. According to this model a continuously distributed trait is manifested as two phenotypes, which are determined by a threshold of sensitivity, individuals below the threshold appearing as one phenotype and individuals above the threshold appearing as the alternate (Falconer, 1989). Now suppose we have some trait such as body size that we know to be normally distributed: we can in principle use the threshold model by arbitrarily assigning a threshold and scoring animals as 0 or 1, depending upon whether they are larger or smaller than the threshold. We shall obtain the same heritability of body size by either using the data directly or in the 0,1 form and using the threshold method. Normally we would adopt the former method, because to classify individuals into

dichotomous categories obviously loses information and thus will generally increase the standard error of the estimate. If, however, the data are not normally distributed and we cannot find a transformation that normalizes the data we can check on the robustness of the estimate of the heritability by assuming that a normalizing transformation exists, dividing the data into two sets (preferably close to equality) and estimating the heritability of the ‘normalized’ trait using the threshold model.

To demonstrate the utility of this approach I present results from a simulation study that considers the threshold model as a transformation for the estimate of both the heritability and the genetic correlation.

**Methods**

*Estimating heritability and genetic correlations by the threshold model*

To estimate heritability using the threshold transformation we proceed in three steps.

**1** Divide the data set into two halves of approximately equal size (which minimizes the standard error) and code one half as zeros and the other as ones.

**2** Estimate heritability on the 0–1 scale ( $h^2_{0,1}$ ) using the same procedure for the particular breeding design as would be used if the data were continuous. The threshold transformation is not suitable for offspring–parent designs because the estimate is biased downwards (Roff, 1997; p. 58).

**3** Estimate heritability on the ‘underlying’ untransformed scale using the formula (Dempster & Lerner, 1950)

$$h^2 = h^2_{0,1} \frac{p(1-p)}{z^2} \tag{1}$$

here  $p$  is the mean proportion in the population and  $z$  is the ordinate on the standardized normal curve that corresponds to a probability  $p$  (e.g. if  $p=0.3$ ). The mean proportion  $p$  is approximated by the number of zeros (or ones) divided by the total number of observations. However, if the number of individuals in a family varies, then a weighted estimate is better: for example, for full sibs use the mean proportion per family.

The standard error of the heritability,  $SE(h^2)$ , is obtained in the same manner,

$$SE(h^2) = SE(h^2_{0,1}) \frac{p(1-p)}{z^2} \tag{2}$$

The genetic correlation is estimated using the 0–1 data and the same procedure for the particular breeding

design as would be used if the data were continuous. At least for full-sib data no correction factor is required (Mercer & Hill, 1984). The standard error can be estimated using the jackknife procedure (Roff & Preziosi, 1994).

*Description of the simulation model*

The model simulated a full-sib breeding design with 100 families of 10 individuals per family. For the first series of runs I examined the efficacy of the threshold model to estimate the heritability. For a given heritability I generated 1000 (100 families  $\times$  10 individuals per family) values of trait  $X$  using the algorithm (Roff & Preziosi, 1994)

$$X_{ij} = a_{x,i} \sqrt{\frac{1}{2} h^2_x} + b_{x,i,j} \sqrt{1 - \frac{1}{2} h^2_x} \tag{3}$$

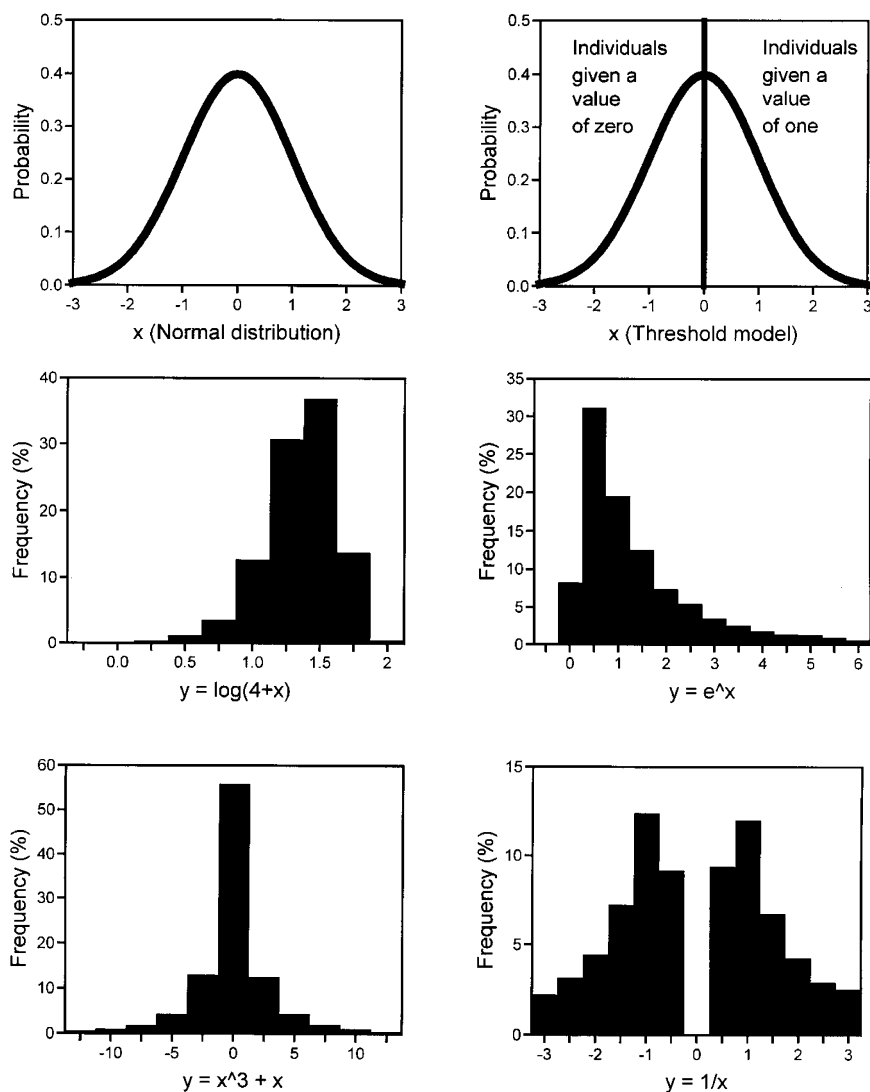
where  $X_{ij}$  is the value of trait  $X$  for individual  $j$  in family  $i$ ;  $a_{x,i}$  is a random normal value,  $N(0,1)$ , common to the  $i$ th family;  $b_{x,i,j}$  is a random normal value,  $N(0,1)$ , for the  $j$ th individual from the  $i$ th family.

Non-normal phenotypic distributions were created by applying some conversion formula  $f(X_{ij})$ . Heritability estimates were then made using the normally distributed values and the untransformed converted phenotypic values. Standard errors were estimated using the formula (Roff, 1997; eqn 2.28)

$$SE(h^2) = 2(1-t)(1+[n-1]t) \left( \frac{2}{n(n-1)(N-1)} \right)^{1/2} \tag{4}$$

where  $t$  is the intraclass correlation coefficient,  $n$  is family size and  $N$  is the number of families.

To compare the above estimates with those obtained using the threshold transformation I divided the original data set into two by designating all values less than zero as 0 and all those above zero as 1 and then proceeding as described in the previous section. Note that the same result would be obtained by dividing the converted data set at its median value: thus regardless of the conversion formula used the threshold transformation gives the same heritability estimate. I examined four non-normal distributions using the conversion formulae: (1)  $f(x) = \log(4+x)$ ; (2)  $f(x) = e^x$ ; (3)  $f(x) = x^3 + x$ ; (4)  $f(x) = 1/x$  (Fig. 1). The logarithmic conversion produced a distribution mildly skewed to the left with a modest kurtosis, whereas the exponential conversion produced a strong right skew and larger kurtosis (Fig. 1, Table 1). Both the polynomial and hyperbolic conversions generated symmetric distributions, the former with a kurtosis approximately midway between the logarithmic and



**Fig. 1** Illustrations of the six distributions used in the simulation study. The four non-normal distributions were generated using 10 000 data points.

	$x$ (normal)	$\log(4 + x)$	$e^x$	$x^3 + x$	$1/x$
Median	-0.012	1.383	0.988	-0.012	-0.382
Mean	-0.005	1.349	1.644	-0.010	0.140
Variance	1.007	0.078	4.492	21.751	5686.549
Skewness (G1)	0.011	-1.206	5.030	-0.146	32.588
Kurtosis (G2)	-0.037	5.129	48.171	23.199	2074.896

**Table 1** Sample statistics for the converted variable,  $f(x)$ , based on 10 000 values

exponential transformations (Fig. 1, Table 1). The hyperbolic was by far the most non-normal distribution being, in fact, bimodal (Fig. 1). For each replicate there were six heritability estimates. For heritability values ranging from 0.1 to 0.9 I generated at each value 10 000 replicates from which I calculated the mean heritability estimate, the mean standard error and the probability

the approximate 95% confidence region ( $= \pm 2$  SE) actually enclosed the true heritability.

To examine the effect of non-normality on genetic correlation estimates and the ability of the threshold transformation to overcome bias I generated for each individual a correlated trait  $Y_{ij}$  using the algorithm (Roff & Preziosi, 1994).

$$\begin{aligned}
Y_{i,j} = & r_G a_{x,i} \sqrt{\frac{1}{2} h_y^2} + a_{y,i} \sqrt{\frac{1}{2} (1 - r_G^2) h_y^2} \\
& + r_E b_{x,i,j} \sqrt{1 - \frac{1}{2} h_y^2} + b_{y,i,j} \sqrt{\left(1 - \frac{1}{2} h_y^2\right) (1 - r_E^2)}
\end{aligned}
\quad (5)$$

where  $a_{x,i}$ ,  $a_{y,i}$  are random normal values,  $N(0,1)$ , common to the  $i$ th family;  $b_{x,i,j}$ ,  $b_{y,i,j}$  are random normal values,  $N(0,1)$ , of the  $j$ th individual from the  $i$ th family;  $h_x^2$ ,  $h_y^2$  are the heritabilities of traits  $X$  and  $Y$ , respectively;  $r_G$  is the genetic correlation between traits  $X$  and  $Y$ ;  $r_E$  is the environmental correlation between traits  $X$  and  $Y$ , given by

$$r_E = \frac{r_P \frac{1}{2} r_E h_x h_y}{\sqrt{1 - \frac{1}{2} h_x^2} \left(1 - \frac{1}{2} h_y^2\right)} \quad (6)$$

where  $r_P$  is the phenotypic correlation.

I examined the following sets of combinations: (1)  $r_P = r_G = 0.5$ ,  $h_x^2 = h_y^2 = 0.1$  to  $0.9$  in increments of  $0.1$ ; (2)  $r_P = r_G = 0.2$ ,  $h_x^2 = h_y^2 = 0.5$ ; (3)  $r_P = r_G = 0.8$ ,  $h_x^2 = h_y^2 = 0.2$ ; and (4)  $r_P = r_G = 0.8$ ,  $h_x^2 = h_y^2 = 0.7$ . In all of these cases I used 10 000 replicates and applied the same conversion function to both traits. I repeated case (1) using all possible model combinations. As a further test I generated 1000 parameter combinations in which  $r_P$  and  $r_G$  were independently drawn from a random uniform distribution between  $-1$  and  $+1$ , subject to the condition that  $-1 < r_E < +1$ ,  $h_x^2$  and  $h_y^2$  were independently drawn from a random uniform distribution between  $0$  and  $1$ , and the conversion functions for  $X$  and  $Y$  drawn independently and with equal probability from the first three conversion functions (the hyperbolic function was not used because, as discussed below, it gave uniformly unacceptable estimates). For each parameter combination I estimated for 100 replicates the mean genetic correlation for the normally distributed data, the threshold-transformed data, and from the non-normal data without transformation. Because of the computer time required to estimate the standard errors, these were computed for 100 of the above random combinations but using 1000 replications in each case.

## Results

### Simulation results

There was excellent agreement between the heritability estimate obtained from the original data and that estimated using the threshold transformation (Fig. 2). The effect of the mild skew introduced by the logarithmic

function had only a small effect on the heritability estimate but the other three non-normal distributions produced underestimates of the true heritability, with the 'hyperbolic' data giving a gross underestimate (Fig. 2). As expected, the average estimated standard error for the threshold transformation was larger than obtained from the normally distributed data (Fig. 2). With the exception of the 'hyperbolic' data, the estimated standard errors of the non-normal data sets were relatively close to that estimated from the normally distributed data. An overall measure that incorporates the bias in the estimate and the variation about the mean is the mean square error, defined as  $MSE = \text{bias}^2 + \text{variance of the estimate}$ . As might be expected from the above results, the logarithmic distribution had only a small effect on the MSE relative to the normal distribution (Fig. 2). The threshold transformation produced a substantially lower MSE than untransformed estimates made from any of the other distributions (Fig. 2: note the log scale).

Because of the downward bias of the non-normal data, the estimated 95% confidence interval was typically considerably smaller than the desired interval (Table 2). Except for the lowest heritability the threshold transformation correctly estimated the 95% confidence interval, whereas there was a small bias towards a confidence interval that was slightly too small in the case of the normally distributed data (Table 2).

The results from the six different analyses of the estimation of the genetic correlation were essentially identical and so here I present only the results for the first ( $r_P = r_G = 0.5$ ,  $h_x^2 = h_y^2 = 0.1$  to  $0.9$  in increments of  $0.1$ , both traits follow the same distribution) and sixth (random assignment of parameter values and models) set of combinations. Considering first the results from case 1. The genetic correlation estimates for the hyperbolic data were extremely poor giving values greater than  $+2$  or less than  $-2$  in approximately 90% of cases. For the other non-normal distributions the genetic correlation estimates underestimated the true value, the underestimate increasing with the heritability (Fig. 3). However, with the exception of the hyperbolic data, this tendency is less than 10% of the estimate (Fig. 3). The mean estimates from the threshold transformation are all within 5% of the correct value.

Because of the very poor estimates obtained from the hyperbolic data the 'random assignment' analysis was restricted to the other three non-normal distributions. In some cases the estimated genetic correlation was inordinately high or low: because such estimates are likely to be discarded and hence not produce incorrect conclusions I dropped all estimates of the genetic correlation greater than  $2$  or less than  $-2$ . The probability of obtaining such discrepant estimates depends upon the

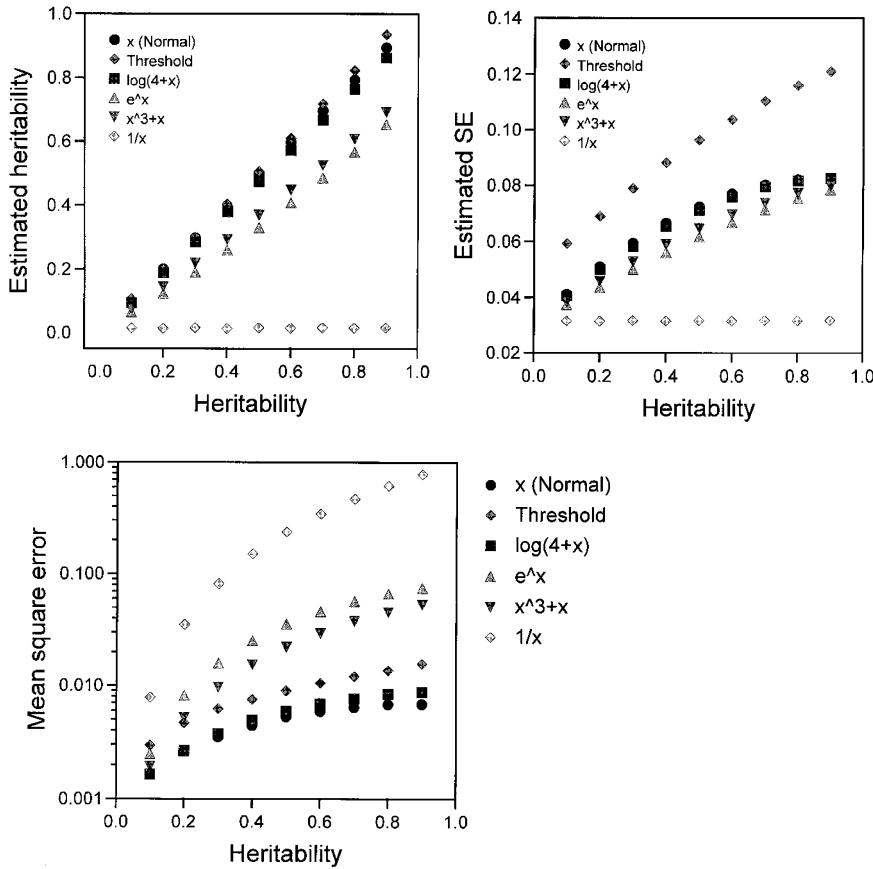


Fig. 2 Top left panel: the mean estimated heritability for the six different data distributions. Top right panel: the mean of the estimated standard error for the above heritability estimates. Bottom panel: mean square error (MSE) for the six distributions.

Table 2 The proportion of cases (determined from 1000 replicates per heritability) in which the estimated 95% confidence intervals included the true heritability

$h^2$	$x$ (normal)	Threshold	log (4 + $x$ )	$e^x$	$x^3 + x$	$1/x$
0.1	0.95	0.98	0.93	0.80	0.87	0.11
0.2	0.94	0.95	0.93	0.55	0.72	0.01
0.3	0.94	0.95	0.92	0.41	0.62	0.00
0.4	0.94	0.95	0.91	0.33	0.53	0.00
0.5	0.94	0.95	0.92	0.29	0.46	0.00
0.6	0.94	0.95	0.91	0.27	0.40	0.00
0.7	0.95	0.95	0.92	0.24	0.35	0.00
0.8	0.95	0.95	0.91	0.23	0.32	0.00
0.9	0.95	0.95	0.91	0.21	0.29	0.00

heritabilities, with geometric mean heritabilities less than approximately 0.3 frequently producing unreasonable estimates (Fig. 4). Estimated genetic correlations from normally distributed data were less likely to produce unacceptable estimates than data transformed using the threshold model (mean number of ‘acceptable’ replicates for normal data = 97.1 vs. 95.5 for threshold-

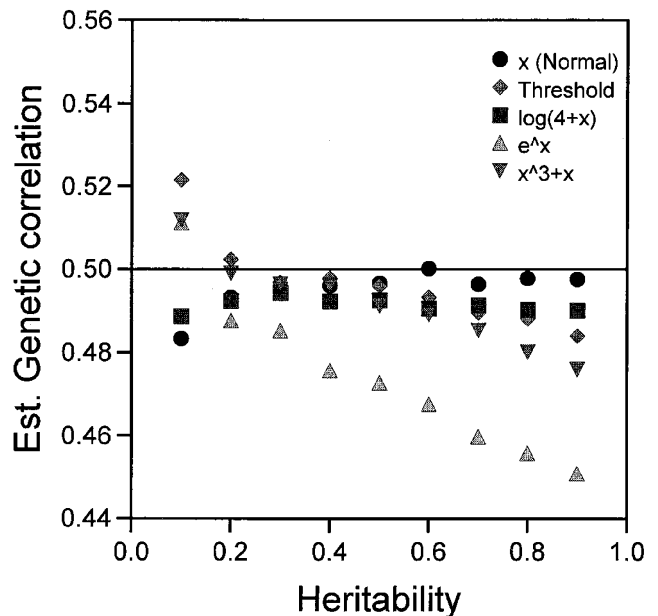
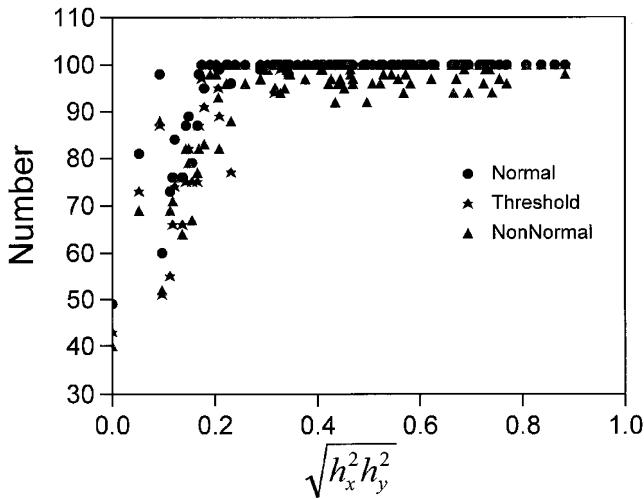


Fig. 3 Mean estimated genetic correlation as a function of the heritability and data distribution. Each mean estimate based on 10 000 replicates.



**Fig. 4** The number of ‘acceptable’ estimates of the genetic correlation as a function of the geometric mean of the heritabilities. The maximum number possible number of ‘acceptable’ estimates per geometric heritability is 100.

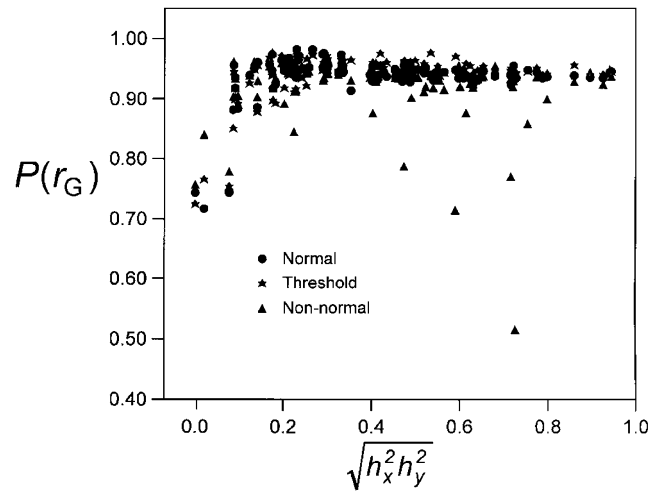
transformed data;  $t = 13.5$ , d.f. = 999,  $P < 0.0005$ ). Similarly the threshold transformation produced more ‘acceptable’ estimates than the non-normal data (mean ‘acceptable’ for the non-normal data = 93.9;  $t = 13.6$ , d.f. = 999,  $P < 0.0005$ ).

There was no evidence for a bias in the genetic correlation estimate for either the threshold transformation ( $= -0.003$ ,  $t = 1.10$ , d.f. = 999,  $P = 0.27$ ) or the untransformed data ( $= -0.003$ ,  $t = 1.27$ , d.f. = 999,  $P = 0.27$ ). The heritability estimates from the normal data were less variable (mean standard deviation of estimate = 0.157) than either the threshold data (0.202) or the non-normal data (0.188). Overall the analyses indicate that the threshold transformation provides a satisfactory estimate of the genetic correlation if there is no transformation that will produce a continuous normal distribution.

The estimated confidence intervals using the non-normal data were smaller than those for either the normal or threshold-transformed data (Table 3). When the

**Table 3** Mean proportion of times,  $P(r_G)$ , that the estimated 95% confidence interval actually enclosed the true value of the genetic correlation. For each of 100 parameter combinations  $P(r_G)$  was estimated from 1000 replicates

Distributions	Mean	SE	Median
Both normal	0.938	0.004	0.944
Both transformed using the threshold model	0.940	0.004	0.948
Both non-normal	0.921	0.006	0.937

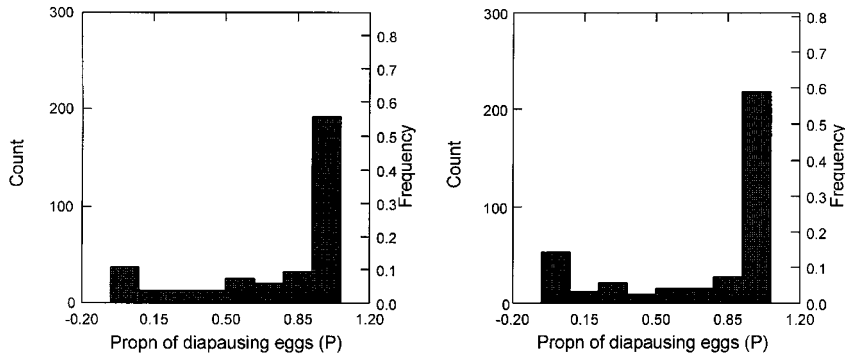


**Fig. 5** The proportion of times the estimated 95% confidence interval of the estimated genetic correlation included the true genetic correlation. Each estimate is based on 1000 replicates.

geometric mean heritability was less than approximately 0.15 all three types of distributions tended to generate confidence intervals that were too small but this propensity was evident in the non-normal distributions for all values of the geometric mean heritability (Fig. 5). There was a significant difference in the proportion of times,  $P(r_G)$ , that the estimated 95% confidence interval enclosed the true genetic correlation ( $F_{2,297} = 4.67$ ,  $P = 0.010$ ). Pairwise comparison using Tukey’s test showed that there was no difference between the normal and threshold-transformed data ( $P = 0.911$ ), but both the normal and transformed data were significantly different from the non-normal data (comparing normal vs. non-normal,  $P = 0.041$ ; comparing threshold-transformed vs. non-normal,  $P = 0.013$ ).

**An example**

To illustrate the technique using a real data set I shall use data on the production of diapause and non-diapause eggs by the cricket *Allonemobius socius* (Roff & Bradford, 2000). In this study egg diapause was considered a trait of the mother and defined as the proportion of diapausing eggs produced by a female of a given age. The proportion of eggs varied according to age and rearing environment. The effects of two environments, corresponding to early and late periods in the summer, and four age groups (days 9–12, 13–16, 17–20, 21–24, post eclosion) were assessed. As can be seen from the two representative distributions shown in Fig. 6, the distribution of diapause proportion was highly non-normal at all ages and in both environments. In the



**Fig. 6** Two examples of the distribution of the proportion of diapausing eggs laid by female *Allonemobius socius*. The left panel shows the distribution for females aged 9–12 days in the late environment, and the right panel for females aged 13–16 in the late environment.

original analysis Roff & Bradford (2000) estimated the heritabilities without transforming the data but assessed the statistical significance by a randomization test.

Approximately 50% of females produced only diapausing eggs at a given age: therefore, to use the threshold transformation I designated the categories as 1 = all diapausing eggs and 0 = some diapausing eggs. In the early environment the heritabilities estimated using the threshold transformation are consistently larger than those of the untransformed data (Table 4). As expected, the standard errors from the threshold transformed data are larger than those of the untransformed data. The randomization test shows that both sets of estimates are highly significant, as do the confidence regions estimated using the standard errors (Table 4). In contrast to the results for the early environment, the two sets of estimates in the late environments are very similar. The standard errors of the threshold estimates are larger than those of the untransformed data and in two cases the 95% confidence region includes zero. The randomization test indicates that in all cases the heritabilities are

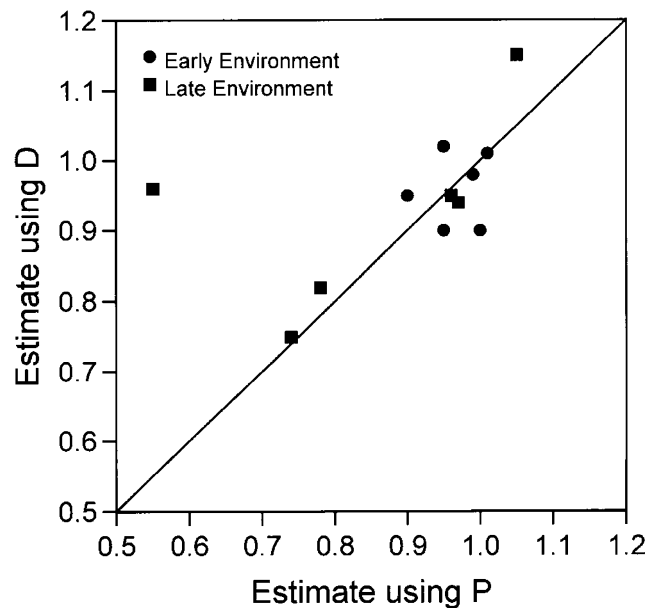
significant (if the probabilities are not Bonferonni adjusted). These results suggest that it is both worthwhile to use the threshold transformation and to apply the randomization test.

According to the simulation analysis the genetic correlation estimate can be estimated directly from the untransformed data. For the *A. socius* data I calculated the genetic correlation between age groups and found, as predicted, that the two sets of estimates corresponded very well (Fig. 7). One set of estimates diverged, the genetic correlation from the untransformed data being 0.55, whereas that from the threshold transformed data was 0.96. The standard errors for both estimates are large (0.39 and 0.30, respectively) and hence the difference is probably not meaningful.

**Table 4** Heritability estimates for the proportion diapausing eggs produced by female *Allonemobius socius* under two rearing conditions and at four ages

Traits†	$h^2$	SE	$P_{rand}‡$
Early environment			
P <sub>1</sub> , D <sub>1</sub>	0.40, 0.50	0.10, 0.16	0.0002, 0.0002
P <sub>2</sub> , D <sub>2</sub>	0.43, 0.78	0.11, 0.19	0.0002, 0.0002
P <sub>3</sub> , D <sub>3</sub>	0.49, 0.67	0.11, 0.17	0.0002, 0.0002
P <sub>4</sub> , D <sub>4</sub>	0.48, 0.73	0.13, 0.19	0.0002, 0.0002
Late environment			
P <sub>1</sub> , D <sub>1</sub>	0.23, 0.22	0.10, 0.17	0.0038, 0.0382
P <sub>2</sub> , D <sub>2</sub>	0.29, 0.32	0.09, 0.14	0.0002, 0.0060
P <sub>3</sub> , D <sub>3</sub>	0.30, 0.42	0.10, 0.18	0.0004, 0.0010
P <sub>4</sub> , D <sub>4</sub>	0.17, 0.36	0.09, 0.20	0.0184, 0.0042

†P<sub>i</sub>, proportion of diapausing eggs at age i. D<sub>i</sub>, proportion of diapausing eggs transformed using the threshold model, where 1 = only diapausing eggs and 0 = some nondiapausing eggs.  
‡Probability obtained from randomization method.



**Fig. 7** Correspondence between the genetic correlation estimated using the untransformed data (P) and that obtained using the threshold transformation (D). The genetic correlations are for proportion of diapausing eggs produced at different ages within the same environment.

## Conclusions

The threshold transformation gave unbiased estimates of the heritability and genetic correlation in all cases. Further, the 95% confidence limits correctly included the true heritability value in the required 95% of cases. The estimated confidence region for the genetic correlation was also correct provided that the geometric mean heritability was greater than approximately 0.15, a restriction that applied also to the normally distributed data (Fig. 5). Thus in the absence of a continuous transformation that produces a normal distribution the threshold transformation is an appropriate method. However, if a continuous transformation is available then this is to be preferred because it will give smaller confidence regions. The method is readily applied to simple breeding designs such as the full- or half-sib method (Roff, 1997; pp 55–57), but difficulties could arise in more complex designs (e.g. designs which include several generations of relatives of different degree). In these cases the approach can be applied using a maximum-likelihood approach (for an example, of its application to a selection experiment see Roff, 1997; pp 143–146).

Even data that are rather badly non-normal (e.g.  $f(x) = e^x$ ) give estimates of heritability and genetic correlation that are reasonably close to the correct value (Figs 1, 3). The problem is that the confidence regions are severely underestimated in the case of the heritability and on average modestly so for the genetic correlation (although gross underestimates can occur, Fig. 5). Thus whereas the point estimates of heritability and genetic correlation are relatively unbiased, the downwards bias

of the estimated confidence region hinders interpretation of any single estimate. Therefore, given data that cannot be transformed it would seem prudent to examine the data using both the raw values and those that have been transformed using the threshold model. If the threshold transformation produces estimates that change the conclusions reached using the raw values then the latter should be viewed with scepticism and the results from the threshold transformation preferred.

## Acknowledgements

This study was supported by an operating grant from the Natural Sciences and Engineering Council of Canada.

## References

- BULMER, M. G. 1985. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, Oxford.
- DEMPSTER, E. R. AND LERNER, I. M. 1950. Heritability of threshold characters. *Genetics*, **35**, 212–236.
- FALCONER, D. S. 1989. *Introduction to Quantitative Genetics*. Longmans, New York.
- MERCER, J. T. AND HILL, W. G. 1984. Estimation of genetic parameters for skeletal defects in broiler chickens. *Heredity*, **53**, 193–203.
- ROFF, D. A. 1997. *Evolutionary Quantitative Genetics*. Chapman & Hall, New York.
- ROFF, D. A. AND BRADFORD, M. R. 2000. A quantitative genetic analysis of phenotypic plasticity of diapause induction in the cricket *Allonemobius socius*. *Heredity*, **84**, 193–200.
- ROFF, D. A. AND PREZIOSI, R. 1994. The estimation of the genetic correlation: the use of the jackknife. *Heredity*, **73**, 544–548.